

# Comparison of ChatGPT and Gemini in Providing Postoperative Information After Tonsillectomy: An Expert-Based Assessment

Karahan Kara<sup>1</sup>

Adana Hospital

Muhammed Dağkiran<sup>2</sup>

[muhammeddagkiran@gmail.com](mailto:muhammeddagkiran@gmail.com)

Adana Hospital

---

## Article

**Keywords:** tonsillectomy, artificial intelligence, large language models, ChatGPT, Gemini, patient education

**Posted Date:** October 22nd, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-10959787/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.  
[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Abstract

Large language models may support postoperative patient education, but their information quality requires evaluation. We compared ChatGPT and Gemini responses to 15 tonsillectomy questions derived from 40 patients and refined by a 10-member expert panel. Turkish responses were generated on August 25, 2026, using GPT-5.6 Sol and Gemini 3.1 Pro. Sixteen blinded otolaryngologists rated scientific accuracy, clinical appropriateness, understandability, and patient safety on five-point scales, providing 1,920 complete ratings. The primary outcome was the four-domain composite. Paired differences were analyzed using a cross-classified mixed-effects model; secondary domain comparisons underwent Holm correction. Composite scores were  $4.35 \pm 0.39$  for ChatGPT and  $4.22 \pm 0.44$  for Gemini (mean difference, 0.138; 95% confidence interval, - 0.184 to 0.459;  $p = 0.381$ ). Accuracy and appropriateness scores were nearly identical. Numerical advantages for ChatGPT in understandability and safety were not statistically significant, and no domain- or question-level comparison was significant after correction. Patient safety received the lowest scores for both models. These findings do not demonstrate superiority or establish equivalence. Lower safety ratings support clinician review before incorporating model-generated information into postoperative education.

## Introduction

Tonsillectomy remains one of the most frequently performed procedures in otolaryngology. Although it is generally undertaken as a short-stay or ambulatory operation, recovery commonly involves substantial throat pain, referred otalgia, dysphagia, reduced oral intake, and the risk of dehydration or postoperative hemorrhage. Because most recovery occurs at home, patients and caregivers must be able to distinguish expected postoperative findings from warning signs and correctly implement recommendations regarding analgesia, hydration, diet, activity, and urgent medical evaluation. Contemporary clinical guidelines therefore identify clear perioperative counseling and evidence-based pain management as central components of safe tonsillectomy care [1,2].

Large language models (LLMs) have emerged as readily accessible tools for supplementing conventional patient education. Their conversational interfaces can provide immediate responses, clarify medical terminology, and adapt information to individual questions. However, LLM outputs are probabilistic, sensitive to model version and prompting, and are not inherently guaranteed to be current, guideline-concordant, or clinically safe. A systematic review of LLMs in preoperative and discharge education found inconsistent effects on patient-centered outcomes and noted that trustworthiness and performance were infrequently evaluated alongside clinical endpoints [3]. Similarly, a systematic review and meta-analysis of generative AI in otolaryngology reported relatively strong performance in educational tasks but substantial between-study heterogeneity, with hallucinations and variable performance persisting in more complex clinical reasoning tasks [4].

Studies focused specifically on tonsillectomy have produced mixed findings. Dhar et al. found that ChatGPT generated thorough responses concerning diet, bleeding, and return to school, but also

produced inaccurate recommendations regarding pain management and follow-up, including a pediatric codeine recommendation [5]. In contrast, Moise et al. reported that 15 of 16 ChatGPT responses regarding indications and post-tonsillectomy complications closely adhered to an established clinical guideline [6]. Wang et al. reported that 34 of 41 (82.9%) guideline-derived ChatGPT responses received the highest accuracy rating, although questions concerning postoperative care and secondary hemorrhage were among the weakest [7]. Polat et al. likewise found that most ChatGPT responses addressing common pediatric otolaryngologic operations were comprehensive or correct, although the material generally exceeded recommended patient reading levels [8]. Abou-Abdallah et al. reported poorer readability and quality for ChatGPT-generated information than for professional ENT patient leaflets and observed that simplification improved readability at the cost of information quality [9]. Bass et al. found that parents rated AI-generated tonsillectomy discharge summaries more favorably than physician-written summaries for perceived medical-information quality and ease of reading, although clinical validity was not directly assessed [10]. Collectively, these findings indicate that fluent presentation does not necessarily ensure completeness, usability, or safety.

Direct comparisons of ChatGPT and Gemini in otolaryngology also suggest that relative performance is task- and domain-dependent. In rhinosinusitis education, ChatGPT-4 achieved a higher rate of correct responses in the treatment category, whereas Gemini produced more inaccurate or incomplete responses in selected categories [11]. Conversely, a guideline-based pediatric otolaryngology study found high accuracy and completeness for both GPT-4 and Gemini without a significant difference in accuracy [12]. Singh et al. likewise found no significant difference between ChatGPT and Gemini across five common pediatric tonsillitis questions ( $p = 0.33$ ), although the outputs from both models exceeded recommended readability levels [13]. In postoperative rhinologic instructions, Shaari et al. found higher understandability for ChatGPT and higher actionability for Gemini [14]. A subsequent comparison of postoperative instructions generated by ChatGPT, Gemini, and DeepSeek found that ChatGPT performed best for readability and Gemini for understandability, but no platform simultaneously met all predefined adequacy thresholds [15]. Thus, results obtained from a single model, version, clinical topic, or evaluation method cannot be assumed to generalize to other settings.

Evidence directly comparing newer-generation ChatGPT and Gemini models for post-tonsillectomy patient information remains limited, particularly when the questions originate from patients and the outputs are assessed by a blinded expert panel across multiple clinically relevant domains. The present study therefore compared responses generated by ChatGPT (GPT-5.6 Sol) and Gemini 3.1 Pro to 15 patient-derived Turkish questions concerning post-tonsillectomy care. The primary objective was to compare the overall composite evaluation of the two models; secondary objectives were to compare scientific accuracy, clinical appropriateness, understandability, and patient safety/misleading-information risk.

## Results

### Expert Characteristics and Data Completeness

All 16 raters confirmed that they were otolaryngologists, provided informed consent, and completed both evaluation forms. Their mean clinical experience was  $14.25 \pm 3.55$  years; the median was 15 years (IQR, 12.00–16.25; range, 7–20 years). Fifteen experts (93.8%) had performed a tonsillectomy during the preceding 12 months. All raters practiced at a training and research hospital (Table 1). The paired datasets contained all 1,920 expected ratings, with no missing or out-of-range values.

Table 1  
Characteristics of the evaluating otolaryngologists

| Characteristic                                                         | Value               |
|------------------------------------------------------------------------|---------------------|
| Number of evaluating otolaryngologists                                 | 16                  |
| Clinical experience, mean $\pm$ SD, years                              | $14.25 \pm 3.55$    |
| Clinical experience, median (IQR), years                               | 15.00 (12.00–16.25) |
| Clinical experience, range, years                                      | 7–20                |
| Tonsillectomy performed in the preceding 12 months, n (%)              | 15 (93.8)           |
| Training and research hospital, n (%)                                  | 16 (100)            |
| Completed both blinded evaluation forms, n (%)                         | 16 (100)            |
| Note. Abbreviations: IQR, interquartile range; SD, standard deviation. |                     |

## Overall and Domain-Specific Model Comparisons

ChatGPT received a numerically higher overall composite score than Gemini ( $4.35 \pm 0.39$  vs  $4.22 \pm 0.44$ ). However, the cross-classified model did not identify a statistically significant overall difference (mean difference, 0.138 points in favor of ChatGPT; 95% CI,  $-0.184$  to  $0.459$ ;  $p = 0.381$ ). The standardized paired effect was small (Cohen’s  $d_z = 0.22$ ). Nine experts gave a higher mean overall score to ChatGPT, whereas seven gave a higher score to Gemini.

Scientific accuracy and clinical appropriateness scores were nearly identical between the models, with mean differences of 0.025 points for both domains. Larger numerical differences favoring ChatGPT were observed for understandability (mean difference, 0.192) and patient safety (mean difference, 0.308); nevertheless, their confidence intervals included zero and neither comparison was statistically significant. None of the four domain-specific comparisons remained significant after Holm correction (all adjusted  $p = 1.000$ ) (Table 2). Both models received their highest mean scores for scientific accuracy and their lowest scores for patient safety.

Table 2  
Comparison of ChatGPT and Gemini ratings

| Outcome                                                                                                                                                                                                                                                                                                                                                                                                                                                               | ChatGPT,<br>mean<br><br>± SD | Gemini,<br>mean<br><br>± SD | Mean<br>difference | 95% CI             | P<br>value | Holm-<br>adjusted<br>P value |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------|-----------------------------|--------------------|--------------------|------------|------------------------------|
| Overall composite                                                                                                                                                                                                                                                                                                                                                                                                                                                     | 4.353 ±<br>0.394             | 4.216 ±<br>0.441            | 0.138              | -0.184<br>to 0.459 | .381       | —                            |
| Scientific accuracy                                                                                                                                                                                                                                                                                                                                                                                                                                                   | 4.600 ±<br>0.340             | 4.575 ±<br>0.379            | 0.025              | -0.263<br>to 0.313 | .858       | 1.000                        |
| Clinical<br>appropriateness                                                                                                                                                                                                                                                                                                                                                                                                                                           | 4.517 ±<br>0.336             | 4.492 ±<br>0.373            | 0.025              | -0.271<br>to 0.321 | .862       | 1.000                        |
| Understandability                                                                                                                                                                                                                                                                                                                                                                                                                                                     | 4.354 ±<br>0.451             | 4.163 ±<br>0.592            | 0.192              | -0.251<br>to 0.635 | .376       | 1.000                        |
| Patient safety                                                                                                                                                                                                                                                                                                                                                                                                                                                        | 3.942 ±<br>0.991             | 3.633 ±<br>1.170            | 0.308              | -0.557<br>to 1.174 | .461       | 1.000                        |
| <p>Note. Scores range from 1 to 5, with higher values indicating more favorable evaluations. Descriptive values are based on the 16 experts' mean ratings. Mean differences are ChatGPT minus Gemini and were estimated using the cross-classified restricted maximum-likelihood model. The overall composite was the primary outcome; Holm correction was applied to the four secondary domains. Abbreviations: CI, confidence interval; SD, standard deviation.</p> |                              |                             |                    |                    |            |                              |

## Robustness and Sensitivity Analyses

The primary finding was unchanged across all robustness analyses. For the overall composite, the expert-level paired t test yielded  $p = 0.384$ , the exact sign-flip test yielded  $p = 0.386$ , and the Wilcoxon signed-rank test yielded  $p = 0.569$ . The 50,000-iteration crossed bootstrap produced a 95% CI of  $-0.172$  to  $0.456$ , and two-way cluster-robust inference yielded  $p = 0.365$ . After excluding the expert who had not performed a tonsillectomy within the preceding 12 months, the overall difference remained nonsignificant (mean difference,  $0.160$ ; 95% CI,  $-0.181$  to  $0.501$ ;  $p = 0.337$ ). These analyses consistently supported a numerical trend toward higher ChatGPT scores without evidence of statistical superiority.

## Inter-rater Reliability and Question-Level Findings

Average-measures absolute-agreement reliability for the overall composite was  $ICC(A,16) = 0.473$  for ChatGPT and  $0.361$  for Gemini. Reliability was highest for scientific accuracy and lowest for patient safety in both models (Table 3). These estimates should be interpreted in the context of the pronounced ceiling distribution in accuracy-related scores and limited between-question variability.

Table 3  
Inter-rater reliability of model evaluations

| Outcome                                                                                                                                                                                                                                              | ChatGPT ICC(A,16) | Gemini ICC(A,16) |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------|------------------|
| Overall composite                                                                                                                                                                                                                                    | 0.473             | 0.361            |
| Scientific accuracy                                                                                                                                                                                                                                  | 0.572             | 0.488            |
| Clinical appropriateness                                                                                                                                                                                                                             | 0.470             | 0.430            |
| Understandability                                                                                                                                                                                                                                    | 0.478             | 0.300            |
| Patient safety                                                                                                                                                                                                                                       | 0.061             | 0.043            |
| <p><i>Note.</i> Values are two-way random-effects, absolute-agreement, average-measures intraclass correlation coefficients based on 16 raters. Overall reliability was calculated using the four-domain composite for each of the 15 questions.</p> |                   |                  |

No individual question showed a statistically significant between-model difference after Holm correction across the 15 exploratory comparisons (Supplementary Table S2). The closest unadjusted finding occurred for the question concerning postoperative analgesic use (Question 14), which favored ChatGPT by 0.250 points (unadjusted  $p = 0.055$ ; Holm-adjusted  $p = 0.820$ ). Gemini received numerically higher scores for Questions 7 and 8, whereas ChatGPT received numerically higher scores for the remaining questions; none of these item-level findings survived correction for multiple testing.

## Discussion

In this blinded, paired expert evaluation, ChatGPT and Gemini both received generally favorable ratings for postoperative tonsillectomy information. ChatGPT had a numerically higher overall composite score, but the estimated difference was small and its confidence interval included zero. Scientific accuracy and clinical appropriateness were nearly identical between the models, and no domain- or question-level comparison remained statistically significant after correction for multiple testing. Accordingly, the present results indicate comparable observed performance under the tested conditions, but they should not be interpreted as establishing formal equivalence between the two models.

The high scientific-accuracy and clinical-appropriateness scores are consistent with reports showing that contemporary LLMs can answer many common otolaryngology questions in a clinically acceptable manner. Moise et al. found close guideline adherence in 15 of 16 ChatGPT responses concerning tonsillectomy [6], while Rossi et al. reported high accuracy and completeness for both GPT-4 and Gemini across guideline-based pediatric otolaryngology questions, without a significant accuracy difference [12]. Singh et al. also reported no significant overall performance difference between ChatGPT and Gemini for common pediatric tonsillitis questions, although the small question set and readability limitations constrain direct comparison with postoperative care [13]. However, other studies have identified clinically important shortcomings. Dhar et al. described inaccurate recommendations concerning analgesia and follow-up in otherwise detailed post-tonsillectomy answers [5], and Becerik et

al. observed model- and category-specific differences when ChatGPT-4 and Gemini were used for rhinosinusitis education [11]. These apparently divergent findings are compatible with a task-dependent pattern: performance varies with the model version, prompt construction, clinical topic, reference standard, and evaluation framework rather than representing a stable property of a platform.

ChatGPT showed larger numerical advantages in understandability and patient safety than in accuracy or clinical appropriateness, although neither difference was statistically significant. Previous comparative evidence similarly suggests that presentation quality is not uniformly linked to factual quality. Shaari et al. found higher understandability for ChatGPT but higher actionability for Gemini in postoperative rhinologic instructions [14]. AlZabin et al. found that ChatGPT performed best for readability whereas Gemini performed best for understandability in postoperative pediatric otorhinolaryngology instructions, with no model satisfying all predefined adequacy thresholds [15]. Polat et al. found that most ChatGPT answers about common otolaryngologic operations were comprehensive or correct but exceeded recommended patient reading levels [8]. Abou-Abdallah et al. likewise reported a trade-off in which simplifying AI-generated information improved readability while reducing information quality [9]. The present study assessed clinician-perceived understandability rather than readability formulas or comprehension by patients; these related constructs should therefore not be treated as interchangeable.

Patient safety was the lowest-scoring domain for both models and showed the greatest numerical separation in favor of ChatGPT. In addition, the question concerning postoperative analgesic use produced the closest unadjusted between-model finding, although it was not significant after multiplicity adjustment. This pattern is clinically important because a response can be broadly accurate yet remain unsafe through omission, ambiguity, insufficient escalation advice, or an inappropriate medication recommendation. This concern is supported by Wang et al., whose 41-question tonsillectomy assessment identified postoperative care and secondary hemorrhage as the two weakest ChatGPT responses despite high overall accuracy [7]. Systematic reviews of LLM-supported perioperative education have emphasized that trustworthiness and safety are less consistently assessed than readability or accuracy [3], while broader otolaryngology evidence continues to document hallucinations and variable performance in complex clinical tasks [4]. High average ratings should therefore not be interpreted as evidence that unsupervised outputs are suitable for direct clinical use.

Inter-rater agreement was modest for the overall composite and particularly low for patient safety. This finding may reflect genuine differences among experts in the threshold used to classify an omission or imprecise statement as a safety concern. It may also be partly attributable to the compressed score distribution: when most questions receive similarly high ratings, limited between-question variance lowers the intraclass correlation even if absolute differences between raters are not large. The higher reliability for scientific accuracy than for patient safety supports the view that factual correctness was judged more consistently than the potential consequences of incomplete or misleading advice. Using multiple raters and accounting for both expert- and question-level clustering were therefore important features of the present analysis.

This study has several strengths. The questions originated from patients rather than being selected solely by investigators and were refined by an experienced bilingual expert panel. Both models received the identical finalized Turkish questions on the same date, the outputs were retained verbatim, and experts evaluated anonymized Response A/B sets. The paired design ensured that every expert assessed both models, while the cross-classified model accounted simultaneously for repeated observations within experts and within questions. The complete dataset, prespecified primary outcome, multiplicity correction, and convergent sensitivity analyses further strengthened the robustness of the findings.

Several limitations should also be considered. This was a single-center study with 16 otolaryngologists and a focused set of 15 Turkish-language questions, which may limit generalizability to other institutions, languages, prompts, and clinical contexts. LLM outputs are version- and time-dependent; the findings describe responses generated on August 25, 2026 and may not apply after model updates. Expert ratings measure perceived information quality rather than patient comprehension, behavioral outcomes, or postoperative complications. Although the paired design improved efficiency, the study was not powered or designed as a formal equivalence or noninferiority trial. Finally, the modest inter-rater agreement, particularly for patient safety, indicates that some domain estimates remain sensitive to differences in expert judgment.

The findings support a role for LLMs as clinician-supervised adjuncts to standardized postoperative education rather than as autonomous substitutes for professional counseling. Bass et al.'s parent-centered comparison suggests that AI-generated tonsillectomy discharge summaries may be acceptable to caregivers; however, the study assessed perceived rather than validated clinical quality [10]. Any implementation should preserve institution-specific instructions, clearly identify urgent warning signs, and require clinical review of medication and escalation advice. Future studies should use multicenter and multilingual samples, include direct patient assessments of comprehension and actionability, compare outputs across repeated model versions, and evaluate whether clinician-edited LLM information improves adherence, anxiety, unscheduled contact, or postoperative outcomes. Safety-focused scoring rubrics with explicit anchors may also improve agreement among evaluators.

ChatGPT and Gemini generated postoperative tonsillectomy information that was rated highly for scientific accuracy and clinical appropriateness, with no statistically significant evidence that either model was superior overall or within individual evaluation domains. ChatGPT showed numerical advantages in understandability and patient safety, but these differences were uncertain and did not remain significant after multiplicity adjustment. The lower and less consistent safety ratings for both platforms reinforce the need for clinician review before LLM-generated content is incorporated into patient education.

## Methods

### Study Design and Setting

The study was approved by the Scientific Research Ethics Committee of Adana City Training and Research Hospital (August 20, 2026; decision no. 1512; meeting no. 28). All methods were performed in accordance with the relevant guidelines and regulations. This single-center, blinded, paired, cross-sectional evaluation study was conducted at Adana City Training and Research Hospital, Türkiye. The study compared responses generated by two large language models to frequently asked questions concerning post-tonsillectomy care. The artificial intelligence-generated responses were evaluated exclusively for research purposes and were not used to guide clinical care. Informed consent was obtained from all study participants, including the patients who contributed questions and the participating otolaryngologists. Written informed consent was obtained from all 40 patients before their questions were collected. All participating otolaryngologists reviewed an electronic information statement and provided voluntary informed consent before completing the evaluation. The analytic dataset contained no patient-level clinical information.

## **Development of the Question Set**

After providing written informed consent, each of the 40 patients attending an otolaryngology outpatient clinic was asked to state the question they most wanted answered about post-tonsillectomy care. The collected questions were used to develop a set of 15 frequently asked questions. Recurrent or semantically overlapping questions were grouped and refined by a 10-member expert panel of otolaryngologists. All panel members were bilingual in Turkish and English, practiced at the same center, and had more than five years of clinical experience following completion of specialist training. The final question set covered postoperative pain and referred otalgia, swallowing and oral intake, dietary recommendations, return to school or work, postoperative bleeding, normal wound appearance, fever, analgesic use, and indications for urgent medical evaluation.

## **Generation of Artificial Intelligence Responses**

On August 25, 2026, the identical set of 15 finalized Turkish questions was submitted separately to ChatGPT (GPT-5.6 Sol; OpenAI) and Gemini (Gemini 3.1 Pro; Google LLC). The responses were recorded verbatim. No response was edited, corrected, shortened, expanded, or otherwise standardized, and no attempt was made to equalize response length between the models. The original Turkish outputs were retained for expert evaluation. English translations were prepared for publication and checked against the original Turkish material; these translated questions and responses are presented in Supplementary Table S1.

## **Blinding and Expert Evaluation**

Before distribution, all model identifiers and other source-revealing information were removed, and the response sets were presented as Response A and Response B. Model identity was not disclosed to the raters during scoring. Sixteen otolaryngologists independently evaluated both response sets using

separate electronic forms. Each expert completed both forms, and the two datasets were paired according to the predefined anonymous participant order used during data collection.

Each response was scored in four domains: scientific accuracy, clinical appropriateness, understandability for patients or caregivers, and patient safety/misleading-information risk. A five-point Likert scale was used for every domain, with higher scores indicating a more favorable evaluation and 5 representing the highest rating. Each expert therefore provided 120 domain ratings (15 questions × 2 models × 4 domains), producing 1,920 individual ratings. Optional free-text comments were not included in the quantitative analyses.

## Outcome Measures

For each expert–question–model combination, an overall composite score was calculated as the arithmetic mean of the four domain ratings. The overall composite was used as the primary outcome. Scientific accuracy, clinical appropriateness, understandability, and patient safety were analyzed as secondary domain-specific outcomes. Exploratory question-level composite scores were calculated by averaging the four domains for each question.

## Statistical Analysis

Data completeness, valid score ranges, and row-by-row pairing of the two model datasets were verified before analysis. Continuous descriptive data are presented as mean ± standard deviation (SD) and, where appropriate, median, interquartile range (IQR), and range. Categorical data are reported as counts and percentages. Descriptive model scores in the main table are based on the 16 experts' mean ratings, thereby retaining the expert as the independent sampling unit.

The primary comparison was performed on paired ChatGPT-minus-Gemini difference scores using a cross-classified linear mixed-effects model fitted by restricted maximum likelihood. Expert and question were entered as crossed random intercepts, thereby accounting simultaneously for repeated ratings within experts and within questions. Degrees of freedom were approximated using the Welch–Satterthwaite method. Results are presented as mean differences with 95% confidence intervals (CIs) and two-sided p values. The overall composite was tested without multiplicity adjustment; the four secondary domain comparisons were adjusted using the Holm procedure.

Robustness was examined using expert-level paired t tests, Wilcoxon signed-rank tests, exact sign-flip tests, two-way cluster-robust inference, and a crossed bootstrap that independently resampled experts and questions over 50,000 iterations. A further sensitivity analysis excluded the single expert who had not performed a tonsillectomy during the preceding 12 months. Exploratory question-level comparisons used exact sign-flip tests across experts with Holm correction across the 15 questions. Inter-rater reliability was assessed separately for each model using two-way random-effects, absolute-agreement, average-measures intraclass correlation coefficients [ICC(A,16)]. Statistical significance was defined as a two-sided p value < 0.05. Analyses were performed using Python 3.12 with NumPy 2.3.5, pandas 2.2.3, and SciPy 1.17.0.

# Use of Generative Artificial Intelligence in Manuscript Preparation

During preparation of this manuscript, ChatGPT was used to assist with language editing and manuscript organization. The authors reviewed and revised the resulting text and take responsibility for the content.

## Declarations

## Ethics and Informed Consent

The study was approved by the Scientific Research Ethics Committee of Adana City Training and Research Hospital (August 20, 2026; decision no. 1512; meeting no. 28). All methods were performed in accordance with the relevant guidelines and regulations. Informed consent was obtained from all study participants, including the patients who contributed questions and the participating otolaryngologists. Written informed consent was obtained from all 40 patients before their questions were collected. All participating otolaryngologists reviewed an electronic information statement and provided voluntary informed consent before completing the evaluation. The artificial intelligence-generated responses were evaluated exclusively for research purposes, and the analytic dataset contained no patient-level clinical information.

## Consent for Publication

Not applicable.

## Competing Interests

The authors declare no competing interests.

## Funding

The authors received no specific funding for this study.

## Author Contribution

Karahan Kara: conceptualization, methodology, investigation, data curation, formal analysis, and writing—original draft. Muhammed Dağkiran: supervision.

# Acknowledgments

The authors thank the participating otolaryngologists for their expert evaluations.

## Data Availability

The de-identified rating data is available from the corresponding author upon reasonable request, subject to institutional requirements.

## References

1. Mitchell, R. B. et al. Clinical Practice Guideline: Tonsillectomy in Children (Update). *Otolaryngol. Head Neck Surg.* **160**, S1-S42 (2019). <https://doi.org/10.1177/0194599818801757>
2. Aldamluji, N., Burgess, A., Pogatzki-Zahn, E., Raeder, J. & Beloeil, H.; PROSPECT Working Group collaborators. PROSPECT guideline for tonsillectomy: systematic review and procedure-specific postoperative pain management recommendations. *Anaesthesia* **76**, 947-961 (2021). <https://doi.org/10.1111/anae.15299>
3. Wang, M., Ma, H. & Piao, M. Effectiveness of large language models in preoperative and discharge education: a systematic review based on an evaluation framework. *npj Digit. Med.* **9**, 122 (2026). <https://doi.org/10.1038/s41746-025-02302-w>
4. Hack, S. et al. Performance of generative AI across ENT tasks: a systematic review and meta-analysis. *Auris Nasus Larynx* **52**, 585-596 (2025). <https://doi.org/10.1016/j.anl.2025.08.010>
5. Dhar, S. et al. The utility and accuracy of ChatGPT in providing post-operative instructions following tonsillectomy: a pilot study. *Int. J. Pediatr. Otorhinolaryngol.* **179**, 111901 (2024). <https://doi.org/10.1016/j.ijporl.2024.111901>
6. Moise, A., Centomo-Bozzo, A., Orishchak, O., Alnoury, M. K. & Daniel, S. J. Can ChatGPT replace an otolaryngologist in guiding parents on tonsillectomy? *Ear Nose Throat J.* (2024). <https://doi.org/10.1177/01455613241230841>
7. Wang, Y., Yu, J. & Zou, J. Evaluation of the theranostic performance of Chat-GPT in tonsillectomy in children. *Asian J. Surg.* **48**, 3630-3631 (2025). <https://doi.org/10.1016/j.asjsur.2024.12.093>
8. Polat, E. et al. Evaluating the accuracy and readability of ChatGPT in providing parental guidance for adenoidectomy, tonsillectomy, and ventilation tube insertion surgery. *Int. J. Pediatr. Otorhinolaryngol.* **181**, 111998 (2024). <https://doi.org/10.1016/j.ijporl.2024.111998>
9. Abou-Abdallah, M. et al. The quality and readability of patient information provided by ChatGPT: can AI reliably explain common ENT operations? *Eur. Arch. Otorhinolaryngol.* **281**, 6147-6153 (2024). <https://doi.org/10.1007/s00405-024-08598-w>
10. Bass, J., Bodimeade, C. & Choudhury, N. A quality improvement project of patient perception of AI-generated discharge summaries: a comparison with doctor-written summaries. *Ann. R. Coll. Surg.*

*Engl.* **107**, 559-562 (2025). <https://doi.org/10.1308/rcsann.2025.0014>

11. Becerik, C., Yildiz, S., Tepe Karaca, C. & Zer Toros, S. Evaluation of the usability of ChatGPT-4 and Google Gemini in patient education about rhinosinusitis. *Clin. Otolaryngol.* **50**, 456-461 (2025). <https://doi.org/10.1111/coa.14273>
12. Rossi, N. A. et al. Comparative analysis of GPT-4 and Google Gemini's consistency with pediatric otolaryngology guidelines. *Int. J. Pediatr. Otorhinolaryngol.* **193**, 112336 (2025). <https://doi.org/10.1016/j.ijporl.2025.112336>
13. Singh, S. et al. Paediatric tonsillitis: evaluation of artificial intelligence generated responses to commonly asked questions. *Aust. J. Otolaryngol.* **9**, 21 (2026). <https://doi.org/10.21037/ajo-25-48>
14. Shaari, A. L. et al. Comparative analysis of artificial intelligence platforms in generating post-operative instructions for rhinologic surgery. *Indian J. Otolaryngol. Head Neck Surg.* **77**, 601-603 (2025). <https://doi.org/10.1007/s12070-024-05161-1>
15. AlZabin, A. et al. Artificial intelligence in pediatric otorhinolaryngology: assessing readability, understandability, and actionability of postoperative instructions. *Am. J. Otolaryngol.* **47**, 104798 (2026). <https://doi.org/10.1016/j.amjoto.2026.104798>

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [TonsillectomySupplementaryInformation.docx](#)